

Precise identification of Cascading alpha satellite higher order repeats (HORs) in T2T-CHM13 assembly of human chromosome 3

Matko Glunčić^{1*}, Ines Vlahović², Marija Rosandić^{3,4}, Vladimir Paar^{1,4}

Affiliations:

¹Faculty of Science, University of Zagreb, Zagreb, Croatia

²Algebra University College, Zagreb, Croatia

³University Hospital Centre Zagreb (Ret.), Zagreb Croatia

⁴Croatian Academy of Sciences and Arts, Zagreb, Croatia.

*Corresponding author. Email: matko@phy.hr

Abstract

Unraveling the intricate centromere structure of human chromosomes holds profound implications, illuminating fundamental genetic mechanisms and potentially advancing comprehension of genetic disorders and therapeutic interventions. From the recently sequenced complete T2T-CHM13 assembly of human chromosome 3, the precise alpha satellite higher-order repeat (HOR) structure is computed using novel high-precision GRM2023 algorithm with Global Repeat Map (GRM) and Monomer distance (MD) diagrams. This study rigorously identified and structurally analyzed alpha satellite HORs within the centromere. The major alpha satellite HOR array in chromosome 3 reveals a novel Cascading HOR, housing 17mer HOR copies with subfragments of periods 15 and 2. Within each row in the cascading HOR, the monomers are of different types, but different rows within the same cascading 17mer HOR contain more than one monomer of the same type. Each canonical 17mer HOR copy comprises 17 monomers yet belonging to 16 different monomer types. Another pronounced 10mer HOR array is of the regular Willard's type. These revelations highlight the complexity within the chromosome 3 centromeric region, accentuating deviations from anticipated highly regular patterns and hinting at profound information encoding and functional potential within the human centromere.

Keywords: T2T-CHM13 assembly, alpha satellites, Higher order repeats HORs, human centromere, GRM2023 algorithm

Introduction

Recent dramatic advances in long-read sequencing, coupled with innovations in reading length and accuracy, have facilitated the generation of complete human chromosome assemblies such as T2T-CHM13 and have provided coverage of previously elusive complex structural variants [1-7]. Until recently, the centromeric region of the human genome has remained largely uncharted, resembling a genomic "black hole" that restricts our ability to study the organization, variation, and function of centromeres. However, recent technological advancements have made it feasible to conduct comprehensive investigations into the structure and function of the complete human genome. This illuminates the rich genetic variation concealed within these formerly inaccessible regions, which may have implications for both health and disease. In particular, it has spurred studies focusing on higher-order repeats. The unexplored variation underscores the necessity for more comprehensive T2T human genome assemblies derived from genetically diverse individuals. In the initial identification of certain higher-order repeats within complete genomic sequences characterizing the human centromeric region, Altemose et al. [4] recently employed a computational method previously introduced by Paar et al. [8].

Studying the very limited sequencing data available in the past, it was discovered over a century ago that human centromeres contain approximately 171 bp alpha satellite repeat monomers, organized into sequences of n monomers, referred to as n mer HORs [9-22]. The divergence between any two monomers within each HOR copy is significant, ranging from ~20% to 40%. However, HOR copies are further organized in tandem, with the divergence between HOR copies typically being less than 5%. Monomers exhibiting less than 5% of mutual divergence are classified as belonging to the same monomer type. Willard and colleagues found that, within each HOR copy, all constituent monomers belong to different monomer types. This distinct pattern, known as Willard's type HORs, has been extensively studied using the limited sequencing data previously available, despite large gaps in the centromeric region [23-35].

In Willard's type n mer HOR arrays, the most common HOR copy with n constituting monomers is referred to as canonical. Copies in the same HOR array that contain inserts or deletions with respect to the canonical HOR copy are known as variants. The identification of HORs within a given genomic sequence presents a highly intricate computational challenge, requiring sensitive approximations. Until recently, this task was hampered by significant limitations in sequencing technology. The GRM algorithm is a unique algorithm for precise identification of detailed HOR copies, both canonical and all its variants for the Willard's type of HORs, which has no repeat of monomer type within any of HOR copies [8, 18, 28, 36-49].

There are various algorithms available for identifying higher-order periodicities within a given genomic sequence (for example [50-58]), owing to the computational complexity of the problem. It is worth noting that the GRM algorithm offers a distinct advantage in enabling precise

determination of HORs, facilitating the complete identification of both length and structure of all HOR copies. This was recognized in Ref. [4], by using the algorithm NTRprism which is, as pointed out by authors of Ref. [4], similar to the GRM method from Ref. [8]. However, one limitation of this approach is its design specificity for Willard's type HORs, characterized by only one monomer of each type in canonical HOR copies.

To address this limitation, we present and implement a novel algorithm termed GRM2023, which represents an enhanced iteration of our prior Global Repeat Map (GRM) algorithm [8, 18, 28]. GRM2023 extends its characterization beyond Willard's type HORs, further focusing on HORs with repeated monomer types within a Canonical HOR copy. We term these extended HORs as Cascading Higher-Order Repeats (Cascading HORs).

It is worth noting that providing a rigorous description of the structural organization of alpha satellite higher-order repeat sequences (HORs) poses a complex challenge, and discrepancies may arise between the results obtained from different methodologies. One notable advantage of the GRM and GRM2023 tools over alternative algorithms lies in their ability to achieve high precision in identifying HOR copies and elucidating their structure. GRM2023 detects peaks corresponding to alpha satellite HORs, as well as additional peaks that represent repeats (subfragments) not arranged in a tandem fashion. Through the utilization of the GRM2023 algorithm, we are able to verify whether these additional peaks indeed correspond to tandem repeats, thus enhancing the accuracy of our analyses.

Recent searches for the list of alpha satellite higher-order repeats (HORs) within the complete T2T-CHM13 genome assembly of human chromosome 3 have yielded varying results, without precise identification of HOR copies. Previous findings, as referenced in Ref. [4], identified 17mer, 10mer, 5mer, and 4mer HORs, while Ref. [4] reported the presence of 17mer, 15mer, and 2mer HORs. In contrast, an earlier Southern blot analysis of human chromosome 3, detailed in Ref. [4], identified two primary Hind III fragments measuring 2.75 kb and 2.4 kb, which co-segregated in different human-hamster cell hybrids. These fragments corresponded approximately to ~16mer and ~14mer HORs, respectively. Additionally, a 650 bp fragment (~4mer HOR) was cloned and found to exhibit high specificity for the chromosome 3 centromere.

In this study, we conducted a precise identification and analysis of alpha satellite HORs using our high-precision GRM2023 algorithm applied to the complete T2T-CHM13 assembly of human chromosome 3. Our analysis revealed the presence of major 17mer and 10mer HORs, with precise identification of HOR copies, and accurate identification of subfragments corresponding to periods 15 and 2.

Results and discussion

GRM (Global Repeat Map) diagram

In the first step, we identify tandemly organized alpha satellite monomers in T2T-CHM13 assembly of human chromosome 3, enumerated in order of appearance in genomic assembly. Utilizing the high-precision GRM2023 algorithm, we calculate the corresponding GRM diagram

for this array of tandemly organized monomers. In this process, HORs are recognized as prominent peaks in the GRM diagram (Fig. 1a). A peak of period n corresponds to $n \times 171$ bp, representing the n mer HOR. The most prominent GRM peaks for T2T-CHM13 assembly of human chromosome 3 correspond to 17mer and 10mer HORs respectively, with approximate frequencies of GRM peaks at ~ 7000 and ~ 4000 , respectively.

The GRM2023 algorithm represents a novel iteration (as detailed in the Methods section) of the Global Repeat Map (GRM) algorithm, previously utilized for the identification of Willard's type HORs, characterized by the absence of repeat monomer types within a single HOR copy [12, 18, 20, 37, 48]. In contrast, the GRM2023 algorithm is adept at discerning not only Willard's type HORs but also extends its capability to identify HORs exhibiting multiple occurrences of the same monomer type within a single HOR copy. These particular HOR instances are referred to as Cascading Higher-Order Repeat (Cascading HORs) copies. Furthermore, the GRM2023 algorithm facilitates the identification of various other types of monomer repeats, such as intra- and inter-HOR-copy monomer repeats or tertiary HOR repeats, which are referred to as subfragments (SF). In the case of T2T-CHM13 assembly of human chromosome 3, notable repeats of SF types are observed at periods 15, 2, 13, and 19, albeit with frequencies an order of magnitude lower than the two predominant peaks at 17 and 10.

MD (Monomer Distance) diagram

The MD (Monomer Distance) diagram displays the relationship between period and monomer enumeration (see Fig. 1b). Each point on the diagram represents a monomer enumeration on the horizontal axis and its corresponding distance to the next monomer of the same type in a sequentially organized monomer sequence, determining both horizontal and vertical coordinates. These points, termed MD-points, form densely distributed horizontal MD-line segments corresponding to a HOR, with the vertical coordinate reflecting the period of the HOR. For a HOR, these MD-points are densely distributed on the line segment, and with the naked eye they resemble a continuous line in the interval corresponding to constituting monomers. The top MD-line segment within an interval of monomer enumeration corresponds to the n mer HOR array, where n represents the period. As seen from MD diagram (Fig. 1b and Table 1), the most prominent MD-line segment corresponds to 17mer HOR. In the case of Cascading HORs, additional parallel MD-line segments within the same interval of monomer enumeration may appear, exhibiting periods smaller than that of the 17mer HOR. These interspersed repeats, occurring within the HOR array—both intra-HOR-copy and inter-HOR-copies—are termed subfragments. As seen from the MD-diagram (Fig. 1b and Table 1), in the case of 17mer HOR the GRM peaks of periods 15 and 2 correspond to subfragments. The sizable MD-line segments of different periods correspond to identified GRM peaks from (Fig. 1a): major peaks 17 and 10, and to less pronounced weak peaks 15, 2, 13, 19, 5 etc. The location of the 17mer and 10mer major HORs on chromosome 3 is shown in ideogram (Fig. 2).

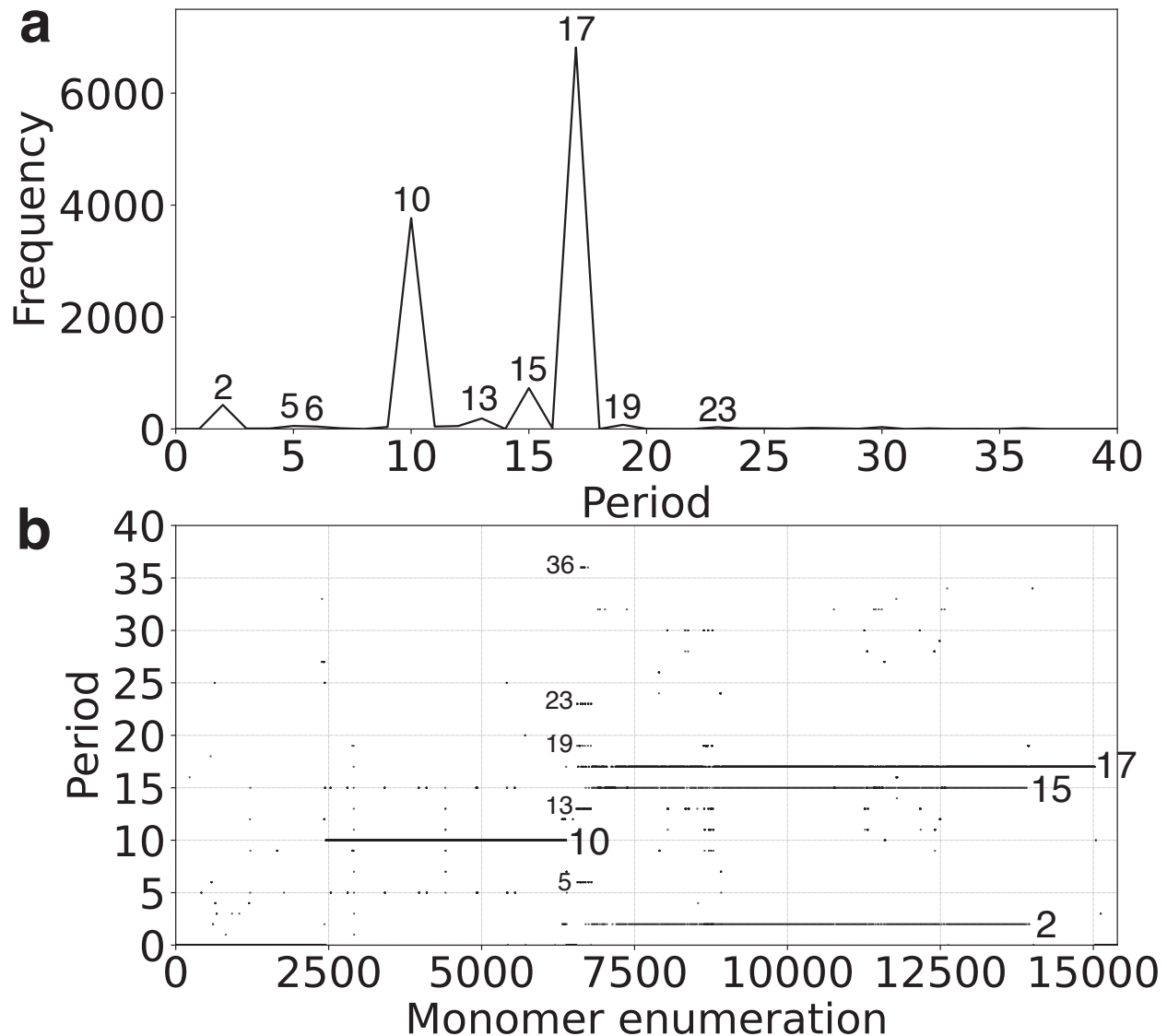


Fig. 1. GRM (Global Repeat Map) diagram and Monomer Distance (MD) diagram for tandemly arranged alpha satellite monomers in complete T2T-CHM13 assembly of human chromosome 3. (a) GRM diagram. Horizontal axis: GRM periods. Vertical axis: frequency of monomer repeats period. Identified major GRM peaks have periods 17, 15, 2, and 10, and minor peaks at 15, 2, 13, 19, 5, 6. The significance of these GRM peaks (HORs or subfragment repeats) can be inferred from the MD diagram. **(b) MD diagram.** Horizontal axis: enumeration of tandemly organized alpha satellite monomers in order of appearance in GRM analysis of T2T assembly. Vertical axis: period (distance between start of a monomer and of the next monomer of the same type). Two pronounced distinct regions with MD-line segments correspond to 17mer HOR (referred to as hor1) and 10mer (referred to as hor2), respectively. The additional MD-line segments at periods 15 and 2 correspond to subsegments of 17mer HOR. In addition, there also some additional weak repeats and sporadic MD-points.

Table 1. Frequency of MD-points for different periods. Number of MD-points for two most frequent periods, 17 and 10, corresponds to the MD-line segments of two major HOR arrays: Cascading 17mer and Willard-type 10mer HOR arrays, respectively. The periods 15 and 2 correspond to subfragments of 17mer HOR. *subfragment denotes relation to a complex repeat in interval of monomer enumeration ~6500-6800 as mentioned in text. The remaining less frequent periods correspond to other less pronounced repeats.

No. of MD points	Period	Repeat pattern
6817	17	Cascading 17mer HOR
3679	10	Willard's type 10mer HOR
731	15	subfragment of Cascading 17mer HOR
430	2	subfragment of Cascading 17mer HOR
188	13	subfragment*
74	19	subfragment*
54	5	subfragment*
52	12	
43	6	
43	11	
36	9	
34	30	
33	23	subfragment*
19	27	
15	16	
14	36	subfragment*

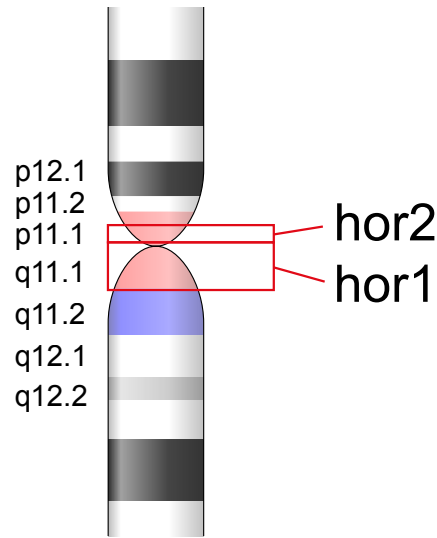


Fig. 2. Ideogram of major alpha satellite HOR arrays in the centromeric region of T2T-CHM13 assembly of human chromosome 3. hor1, Cascading 17mer HOR array; hor2, Willard's type 10mer HOR array.

Aligned scheme for Cascading 17mer HOR array with 15mer and 2mer subfragments

As inferred from the graphical representation provided by the GRM and MD diagrams (Fig. 1a,b), the largest array of higher-order repeats (HOR) within human chromosome 3 is identified as the Cascading 17mer HOR, spanning the genomic interval from 91,779,888 bp to 96,415,046 bp in the T2T-CHM13 assembly. The comprehensive alignment pattern of the Cascading 17mer HOR array, computed using the GRM2023 algorithm, is depicted in Supplementary Fig. S1. Additionally, the predominant constituent of this array, namely the canonical 17mer HOR, is visually depicted through a linear arrangement of its constituent 17 monomers (Fig. 3a).

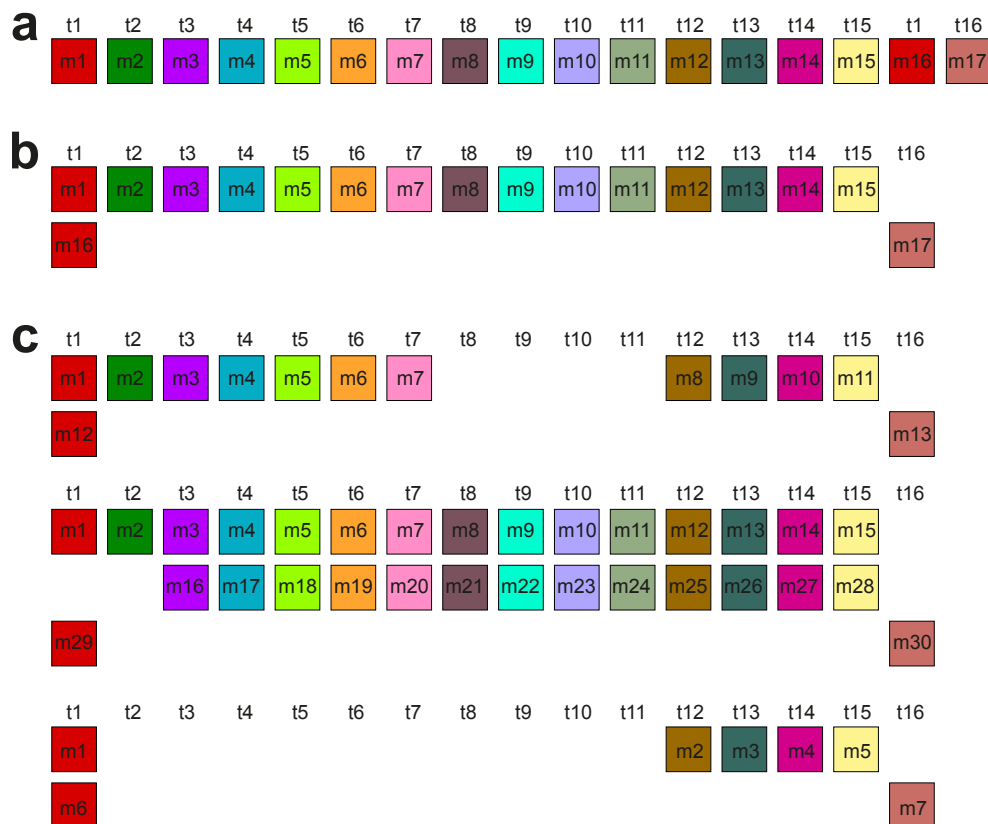


Fig. 3 Aligned scheme of Cascading 17mer HOR canonical HOR copies and some variants
(a) 17mer Canonical HOR copy constituted of 17 monomers (denoted m1,... m17) of 16 different types (t1,... t16) presented in the linear monomeric scheme. The number of different types of monomers in canonical HOR copy is denoted by τ . Each monomer is presented by a colored box.
(b) Cascading Aligned scheme of the canonical 17mer HOR ($n = 17$, $\tau = 16$) corresponding to the linearized scheme in Fig. 3a. Two monomers of the same type are aligned in the first column: monomer m1 of the type t1 in the first row and monomer m16 of the same type t1 in the second row.
(c) Several examples of variant Cascading HOR copies from Supplementary Fig. S1: 13mer, 30mer and 7mer with respect to 17mer HOR array.

Monomers within the 17mer HOR copy, labeled m1 through m17 in order of appearance within canonical HOR copy, are arranged sequentially in a linear fashion, each represented by a distinct colored box. Above each box stands its corresponding type, labeled as t1, t2, and so forth. Different monomer types are distinguished by varying box colors, while monomers of the same type share identical coloring.

The two 17mer cascading HOR monomers, m1 and m16, are classified under the same type, denoted as t1. In instances where the canonical copy exhibits a repetition of monomer types, the linear presentation of the HOR copy is transformed into a cascading format, resulting in a multi-

row arrangement. Each row consists of monomers of distinct types, aligned vertically according to their respective types.

The two 17mer cascading HOR monomers, m1 and m16, are classified under the same type, denoted as t1. In instances where the canonical copy exhibits a repetition of monomer types, the linear presentation of the HOR copy is transformed into a cascading format, resulting in a multi-row arrangement. Each row consists of monomers of distinct types, aligned vertically according to their respective types.

Consequently, the linear single-row depiction of the 17mer canonical HOR copy (Fig. 3a) undergoes transformation into a two-row representation as depicted in Fig. 3b. The first row comprises a linear sequence of monomers, m1 through m15, corresponding to types t1 through t15, respectively. The second row features only two monomers: m16, type t1, aligned with m1 of the same type from the first row, and m17, designated as type t16, positioned to the right of m15 in the first row. This presentation, characterized by aligned monomers based on their types, is termed Cascading 17mer HOR (Fig. 3b).

It's important to note that variants involving adjacent rows, such as (t1, t16) and (t1, t16), exemplified by the 30mer variant in Fig. 3c, also contribute to subfragments of period 2 as a consequence of tertiary HOR. Select segments of the array of Cascading 17mer HOR copies from Supplementary Fig. S1 are depicted in Fig. 4.

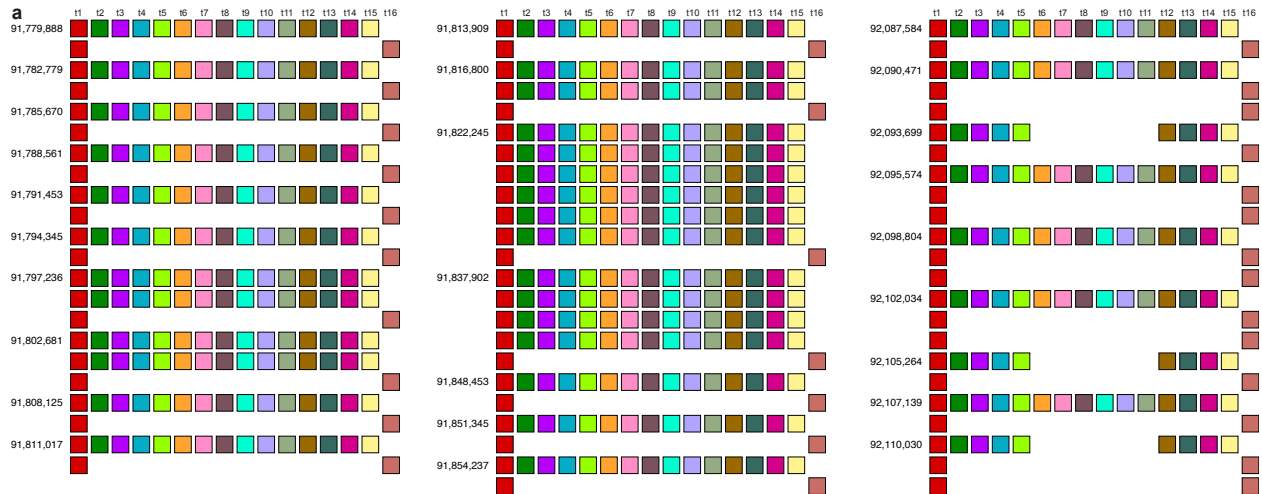


Fig. 4 Aligned scheme of some segments from Cascading 17mer HOR array (a) Segment of the first ten Cascading 17mer HOR copies from position 91,779,888 to 91,811,017. To each HOR copy correspond cascading rows of monomers. The No.1 HOR copy is canonical, consisting of two cascading rows: the first row with 15 monomers of types t1-t15 and the second row with two monomers of types t1 and t16. The next five HOR copies No. 2–6 are of the same canonical structure. The cascading HOR copy No. 7, starting at position 91,797,236, is variant HOR consisting of $15+15+2=32$ monomers (three cascading rows of 15, 15 and 2 monomers,

respectively). This variant HOR copy arises from the canonical HOR copy by duplicating its first row. The next HOR copy No.8 starting at position 91,802,681 is the same as HOR copy No. 7. The next two HOR copies, No.9 and 10, are Canonical (15+2). **(b) Segment of Cascading 17mer HOR copies from position 91,813,909 to 91,854,237.** This segment starts with canonical 17mer HOR copy (15+2 monomers). Then follows an extended HOR copy of $2 \times 15 + 2 = 32$ monomers, which can arise from canonical 17mer HOR copy by a multiplication of the first row in canonical HOR copy. Follows an extended HOR copy of $6 \times 15 + 2 = 92$ monomers, which can arise from canonical 17mer HOR copy by a multiple multiplication of the first row in canonical HOR copy. Further follows a variant HOR copy of $4 \times 15 + 2 = 62$ monomers, which arises from canonical 17mer HOR copy by a multiplication of the first row in canonical HOR copy. After that follows a sequence of canonical 17mer HOR copies. **(c) Segment of Cascading 17mer HOR copies from position 92,087,584 to 92,110,030, giving rise to tertiary period 2 subfragment.** This graphical presentation is also presented in Table 3. The sub-tandem of (t1, t16) doublets within HOR copies gives rise to subrepeats ... t2 t15 t2 t15 ... which due to distances t2-t2 and t15-t15 of 2×171 bp generates intra-HOR tertiary periodicity 2.

It is possible to inspect the accompanying subfragments considering the types of monomers in canonical 17mer HOR copy. The 17 monomers m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13, m14 m15 m16 m17 in a canonical 17mer HOR copy have the corresponding monomer types t1 t2 t3 t4 t5 t6 t7 t8 t9 t10 t11 t12 t13 t14 t15 t16 which for simplicity we write 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 1 16. Analogously, the monomer types in the corresponding neighboring canonical 17mer HOR copy is denoted 1' 2' 3' 4' 5' 6' 7' 8' 9' 10' 11' 12' 13' 14' 15' 1' 16'. Let us consider the two neighboring canonical 17mer HOR copies:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 1 16 1' 2' 3' 4' 5' 6' 7' 8' 9' 10' 11' 12' 13' 14' 15' 1' 16'.

Within the first HOR copy the distance d between the start of two monomers of identical type $t = 1$: between the monomer denoted m1 and the monomer denoted m16 in the initial m-sequence. This distance is equal to 15 units of monomer lengths, i.e., equal to sum of lengths of monomers m1, m2, ... m15, $d = 15$. This is the characteristic intra-HOR-copy distance within each 17mer canonical HOR copy and gives rise to the MD-line segment of period 15 in the MD-diagram. It is referred to as period 15 subfragment. For tandems of canonical 17mer HORs this pattern is equidistant.

Furthermore, the inter-HOR-copy-distance between the second monomer of type 1 in the first HOR copy and the first monomer of type 1 (denoted 1') in the second HOR copy:

... 15 1 16 1' 2' ...

is equal to the sum of lengths of monomers of type 1 and of type 16: $d = 2$. On this principle we obtain two MD-line-segments at periods 15 and 2, referred to as subsegments. They are positioned in the same interval of monomer enumeration as the line-segment corresponding to the 17mer HOR (Fig. 1b). The onset of period 2 arises also due to contributions from variants of 17mer HOR copy, involving tandem repeats of t1 t16 doublets within HOR copies (Fig. 4c). A segment of the Cascading 17mer HOR contributing to period 2 repeats is provided in Complementary Table 2.

Table 2. A segment from Complementary Table of Cascading 17mer HOR contributing to period 2 repeats.

Monomer type	Repeat pattern
t1-t15 t1, t16 t1, t16	Variant15+2+2
t1-t 5, t12-t15 t1, t16	Variant (6+4)+2
t1-t15 t1, t6 t1, t6	Variant15+2+2
t1-t15 t1, t6 t1, t6	Variant15+2+2
t1-t15 t1, t6 t1, t6	Variant15+2+2
t1-t6, t12-t15 t1, t16	Variant (6+4)+2
t1-t15 t1, t16	Canonical 15+2
t1-t5, t12-t15 t1, t16 t1, t16	Variant
t1-t5, t12-t15 t1, t16 t1, t16	Variant
t1-t5, t12-t15 t1, t16 t1, t16	Variant
t1-t15 t1, t16	Canonical 15+2

Furthermore, within a specific range of monomer enumeration, spanning from ~6500 to ~6800, a highly intricate repeating pattern emerges, comprised of subfragments with periods of 23, 19, 17, 13, 6, and occasionally a less pronounced 36.

Fig. 5 illustrates all HORs in this region, with box colors and monomer type labels consistent with those of the 17mer HOR shown in Figs. 3 and 4. As discerned from Fig. 5, there exist five canonical copies of the Cascade 36mer HOR, predominantly composed of the same monomers as the 17mer HOR. Each canonical 36mer HOR includes 16 distinct types of monomers present in the 17mer

HOR (t1 to t16) along with two additional monomer types, t17 and t18. These monomers are largely arranged in the canonical 36mer HOR in the same sequence as in the canonical 17mer HOR, except for the insertion of monomers t16 and t17 between t1 and t2. Furthermore, the canonical 36mer HOR is characterized by a significant number of monomer duplications, with each individual HOR unit containing three copies of t1, t2, t3, t4, t16, and t17, as well as two copies of t5, t11, t12, t13, t14, and t15. Thus, from only 18 distinct monomer types, a 36mer HOR is formed, resulting in a large number of subfragments in Figure 1b. Following the final variant copy of this HOR, commencing at position 91,778,533, the 17mer HOR continues.

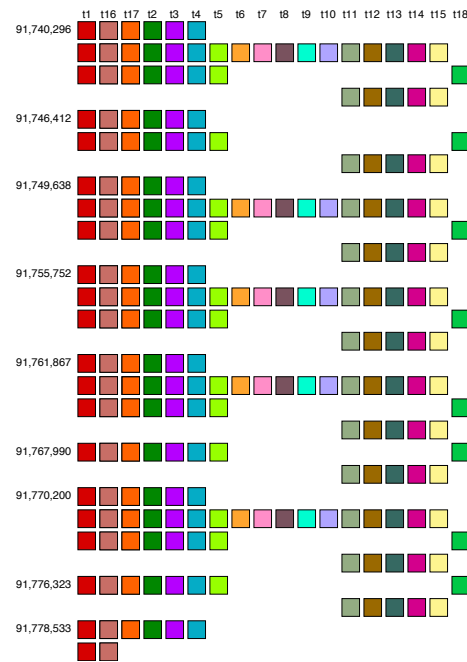


Fig. 6 Aligned scheme of the entire array of Cascading 36mer alpha satellite HOR. The number on the left side indicate the initial position of the first monomer in each row of HOR copy. The box colors and monomer type labels align consistently with those of the 17mer HOR. There are five canonical copies of the Cascade 36mer HOR, primarily composed of the same monomers found in the 17mer HOR. To each HOR copy correspond cascading rows of monomers. The HOR copies No 1, 3, 4, 5 and 7 are canonical, consisting of four cascading rows: the first row with 6 monomers of types t1, t16, t17, t2-t4, the second row with 17 monomers of types t1, t16, t17, t2-t15, the third row with 8 monomers of types t1, t16, t17, t2-t5, t18, and forth row with 5 monomers of types t11-t15 ($6+17+8+5=36$). Between these canonical HOR copies, four variant HOR copies with significantly fewer monomers in each HOR unit are dispersed.

Aligned scheme for Willard's type alpha satellite 10mer HOR array

As observed in the MD diagram (Fig. 1b), the 10mer HOR array, designated as hor2, is situated within the monomer enumeration interval between ~2500 and ~6400 determined from T2T-CHM13 assembly. The aligned 10mer HOR scheme for this 10mer HOR array is presented in Supplementary Fig. S2, and the consensus HOR is displayed in Supplementary Table S2. Specific

segments from the aligned 10mer HOR scheme are graphically presented in Fig. 6. The composition of HOR copies in 10mer HOR array from Supplementary Fig. S2 is analyzed in Supplementary Table S3.

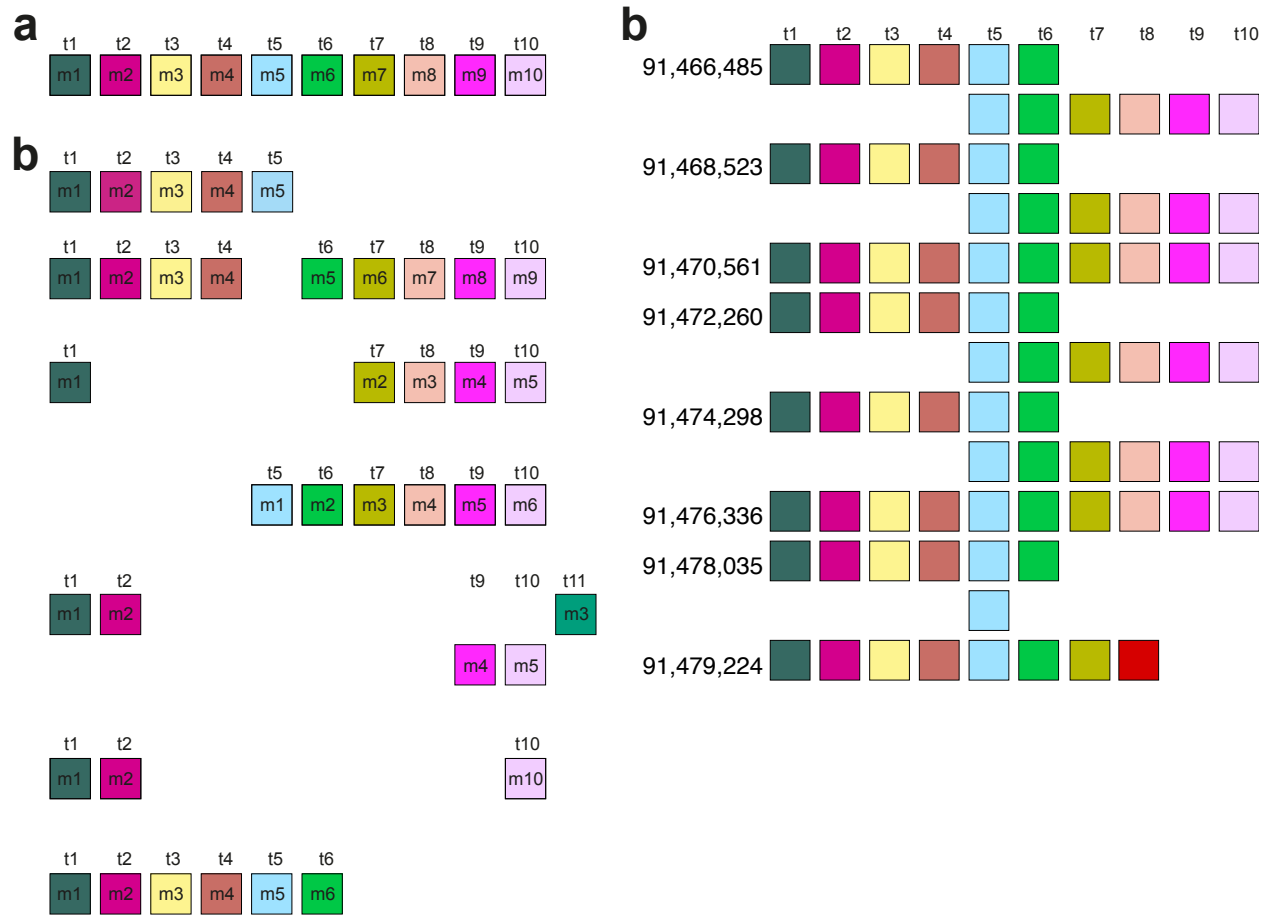


Fig. 6 Aligned scheme of some segments from Willard's type 10mer HOR array. (a)Scheme of canonical 10mer HOR copy. (b) Scheme of several variants in 10mer HOR array. (c)Cluster of canonical and variant HOR copies (2 canonical and 11 variant) at the end of 10mer HOR array

The high percentage of 94% are canonical HOR copies. Variant HOR copies show a strong tendency of clustering in large groups of 76%, 55%, 50%, 50%, and 47%, scattered between large groups of canonical HOR copies, but the composition of monomer types in variants are far from random. Among the monomer types the most frequent in variants are t-t5, t-t6, and t5-t10. Half of variants are located near the end of 10mer HOR array, as transitional region in dissolving the HOR regularity.

Conclusion

Recent advancements in long-read sequencing technologies have led to the achievement of complete human chromosome assemblies, encompassing even the centromere region, for the first time. Consequently, larger variants within the centromere are gradually becoming accessible. Incorporating these findings is crucial for comprehending genome biology and function, given the common association of structural variations with conditions such as cancer, developmental disorders, and complex diseases. The ability to identify structural variations with greater confidence is emerging [1-7].

While some human chromosomes consist predominantly of canonical higher-order repeat (HOR) units, others, such as chromosome 15, harbor numerous types of structural variant HORs. To annotate repeat structure within assembled satellite DNA arrays, the NTRprism algorithm was developed to discover and visualize satellite repeat and HOR periodicity [4], similar to the previous GRM algorithm [8] developed to discover and visualize alpha satellite arrays.

Utilizing the recently sequenced complete T2T-CHM13 assembly of human chromosome 3, the precise alpha satellite Cascading HOR structure is delineated employing our innovative high-precision GRM2023 algorithm with Global Repeat Map (GRM) and Monomer Distance (MD) diagrams. This study rigorously identified and structurally analyzed alpha satellite HORs within the centromere. Notably, the major alpha satellite HOR array in chromosome 3 reveals the novel Cascading 17mer HOR.

Methods

GRM 2023 algorithm.

The alpha satellite HORs were identified in the human chromosome 15 T2T-CHM13 genomic assembly using the GRM2023 algorithm [18, 41, 42]. The GRM2023 algorithm is an efficient and robust method specifically designed to detect and analyze very large repeat units, such as HORs, within genomic sequences. The GRM method generates a global repeat map in a GRM diagram, identifying all prominent repeats in a given sequence without any prior knowledge of the repeats. Furthermore, once the consensus repeat unit is determined using GRM2023, it can be further combined with a search for dispersed HOR copies or individual constituting monomers.

Specifically, alpha satellite HORs in this study were identified through the following steps:

- (i) Using GRMapp (the GRM graphical user interface application is freely available at <http://genom.hazu.hr/tools.html>), alpha satellite monomers were identified within the entire human chromosome T2T-CHM13 assembly. GRMapp provides all tandem repeats (TRs) in the analyzed assembly as its output. From the list of all TRs, those with lengths of ~171 bp were selected and subjected to GRM diagram analysis within GRMapp. To be classified as alpha satellite monomers, the GRM diagram must exhibit peaks at ~171 bp and multiples at ~342 bp ~513 kb, and so on.

- (ii) The extracted alpha satellite monomers were compared to each other, and a divergence matrix was created. From the divergence matrix, monomer families were identified, encompassing all monomers that differ from each other by less than 5%.
- (iii) For each monomer family, a consensus sequence was generated using the stand-alone tool for multiple-sequence alignment, pyabPOA (pyabpoa 1.0.0a0), available at <https://github.com/yangao07/abpoa>. The consensus sequences for all alpha satellite monomer families are provided in Supplementary Tables S5-S8.
- (iv) Chromosome 3 T2T-CHM13 assembly was searched with all consensus sequences using the Edlib open-source C/C++ library for exact pairwise sequence alignment(50). The search was conducted base by base for the entire chromosome, considering both the direct and reverse complement consensus sequences.
- (v) The results of the search in step (iv) were presented graphically (Fig. 3-5) in a way that all monomers of the same family are located in the same column and colored with the same color.

We should like to note that the NTRprism code [4] corresponds to the early version of GRM code and the NTRprism spectrum corresponds to the GRM diagram [18, 41, 42]. In the updated version of GRM used here, the GRM2023 code is extended to also identify the cascading HORs and interspersed HORs.

Acknowledgments: The authors thank to Karen Miga for information on T2T-CHM13 genome assembly.

Funding:

QuantiXLie Centre of Excellence, a project cofinanced by the Croatian Government and European Union through the European Regional Development Fund—the Competitiveness and Cohesion Operational Programme (Grant KK.01.1.1.01.0004).

The grant IP-2019-04- 2757 from Croatian Science Foundation.

Author contributions:

Conceptualization: VP, MG

Methodology: VP, MG

Investigation: MG, IV, MR

Visualization: MG, IV

Funding acquisition: MG, VP

Project administration: MG, VP

Supervision: VP

Writing – original draft: VP, MG

Writing – review & editing: VP, MG, MR

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: Genomic sequence are freely available at the National Center for Biotechnology Information (NCBI) website <https://www.ncbi.nlm.nih.gov>. The GRM graphical user interface application (JAR file) is freely available at our project's website <http://genom.hazu.hr/tools.html>. It can be run on any platform which have Java Runtime Environment (JRE) installed (freely available at <https://www.oracle.com/java/technologies/javase-downloads.html>).

Supplementary Materials

Materials and Methods

Figs. S1 to S2

Tables S1 to S3

References

1. Miga KH: **Centromere studies in the era of 'telomere-to-telomere' genomics.** *Exp Cell Res* 2020, **394**(2):112127.
2. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A et al: **The complete sequence of a human genome.** *Science* 2022, **376**(6588):44-53.
3. Cechova M, Miga KH: **Comprehensive variant discovery in the era of complete human reference genomes.** *Nat Methods* 2023, **20**(1):17-19.
4. Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ, Uralsky L, Ryabov FD, Shew CJ et al: **Complete genomic and epigenetic maps of human centromeres.** *Science* 2022, **376**(6588):eabl4178.
5. Miga KH: **The Promises and Challenges of Genomic Studies of Human Centromeres.** *Prog Mol Subcell Biol* 2017, **56**:285-304.
6. Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, Jain M, Shumate A, Razaghi R, Koren S et al: **Epigenetic patterns in a complete human genome.** *Science* 2022, **376**(6588):eabj5089.
7. Altemose N: **A classical revival: Human satellite DNAs enter the genomics era.** *Semin Cell Dev Biol* 2022, **128**:2-14.
8. Paar V, Basar I, Rosandic M, Gluncic M: **Consensus higher order repeats and frequency of string distributions in human genome.** *Curr Genomics* 2007, **8**(2):93-111.
9. Manuelidis L: **Chromosomal localization of complex and simple repeated human DNAs.** *Chromosoma* 1978, **66**(1):23-32.
10. Wu JC, Manuelidis L: **Sequence definition and organization of a human repeated DNA.** *J Mol Biol* 1980, **142**(3):363-386.
11. Willard HF: **Chromosome-specific organization of human alpha satellite DNA.** *Am J Hum Genet* 1985, **37**(3):524-532.
12. Wayne JS, Willard HF: **Structure, organization, and sequence of alpha satellite DNA from human chromosome 17: evidence for evolution by unequal crossing-over and an ancestral pentamer repeat shared with the human X chromosome.** *Mol Cell Biol* 1986, **6**(9):3156-3165.
13. Willard HF, Wayne JS: **Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat.** *J Mol Evol* 1987, **25**(3):207-214.

14. Waye JS, Willard HF: **Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes.** *Nucleic Acids Res* 1987, **15**(18):7549-7569.
15. Jorgensen AL, Bostock CJ, Bak AL: **Chromosome-specific subfamilies within human alphoid repetitive DNA.** *J Mol Biol* 1986, **187**(2):185-196.
16. Willard HF: **Evolution of alpha satellite.** *Curr Opin Genet Dev* 1991, **1**(4):509-514.
17. Choo KH, Vissel B, Nagy A, Earle E, Kalitsis P: **A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence.** *Nucleic Acids Res* 1991, **19**(6):1179-1182.
18. Gluncic M, Paar V: **Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm.** *Nucleic Acids Res* 2013, **41**(1):e17.
19. Romanova LY, Deriagin GV, Mashkova TD, Tumeneva IG, Mushegian AR, Kisselev LL, Alexandrov IA: **Evidence for selection in evolution of alpha satellite DNA: the central role of CENP-B/pJ alpha binding region.** *J Mol Biol* 1996, **261**(3):334-340.
20. Warburton PE, Willard HF: **Evolution of centromeric alpha satellite DNA: molecular organisation within and between human primate chromosomes.** In: *Human Genome Evolution*. BIOS Scientific Publisher; 1996: 121-145.
21. O'Keefe CL, Matera AG: **Alpha satellite DNA variant-specific oligoprobes differing by a single base can distinguish chromosome 15 homologs.** *Genome Res* 2000, **10**(9):1342-1350.
22. Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, Yurov Y: **Alpha-satellite DNA of primates: old and new families.** *Chromosoma* 2001, **110**(4):253-266.
23. Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF: **Genomic and genetic definition of a functional human centromere.** *Science* 2001, **294**(5540):109-115.
24. Alkan C, Eichler EE, Bailey JA, Sahinalp SC, Tuzun E: **The role of unequal crossover in alpha-satellite DNA evolution: a computational analysis.** *J Comput Biol* 2004, **11**(5):933-944.
25. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Rebase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110**(1-4):462-467.
26. Rudd MK, Wray GA, Willard HF: **The evolutionary dynamics of alpha-satellite.** *Genome Res* 2006, **16**(1):88-96.
27. Alkan C, Ventura M, Archidiacono N, Rocchi M, Sahinalp SC, Eichler EE: **Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data.** *PLoS Comput Biol* 2007, **3**(9):1807-1818.
28. Paar V, Gluncic M, Rosandic M, Basar I, Vlahovic I: **Intragenic higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees.** *Mol Biol Evol* 2011, **28**(6):1877-1892.
29. Hayden KE, Strome ED, Merrett SL, Lee HR, Rudd MK, Willard HF: **Sequences associated with centromere competency in the human genome.** *Mol Cell Biol* 2013, **33**(4):763-772.
30. Terada S, Hirai Y, Hirai H, Koga A: **Higher-order repeat structure in alpha satellite DNA is an attribute of hominoids rather than hominids.** *J Hum Genet* 2013, **58**(11):752-754.
31. Aldrup-Macdonald ME, Sullivan BA: **The past, present, and future of human centromere genomics.** *Genes (Basel)* 2014, **5**(1):33-50.
32. Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ: **Centromere reference models for human chromosomes X and Y satellite arrays.** *Genome Res* 2014, **24**(4):697-707.

33. Shepelev VA, Uralsky LI, Alexandrov AA, Yurov YB, Rogaev EI, Alexandrov IA: **Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly.** *Genom Data* 2015, **5**:139-146.
34. Sullivan LL, Chew K, Sullivan BA: **alpha satellite DNA variation and function of the human centromere.** *Nucleus* 2017, **8**(4):331-339.
35. Uralsky LI, Shepelev VA, Alexandrov AA, Yurov YB, Rogaev EI, Alexandrov IA: **Classification and monomer-by-monomer annotation dataset of suprachromosomal family 1 alpha satellite higher-order repeats in hg38 human genome assembly.** *Data Brief* 2019, **24**:103708.
36. Rosandic M, Paar V, Basar I: **Key-string segmentation algorithm and higher-order repeat 16mer (54 copies) in human alpha satellite DNA in chromosome 7.** *J Theor Biol* 2003, **221**(1):29-37.
37. Rosandic M, Paar V, Basar I, Gluncic M, Pavin N, Pilas I: **CENP-B box and pJalpha sequence distribution in human alpha satellite higher-order repeats (HOR).** *Chromosome Res* 2006, **14**(7):735-753.
38. Rosandic M, Paar V, Gluncic M, Basar I, Pavin N: **Key-string algorithm--novel approach to computational analysis of repetitive sequences in human centromeric DNA.** *Croat Med J* 2003, **44**(4):386-406.
39. Rosandic M, Gluncic M, Paar V, Basar I: **The role of alphoid higher order repeats (HORs) in the centromere folding.** *J Theor Biol* 2008, **254**(3):555-560.
40. Paar V, Pavin N, Rosandic M, Gluncic M, Basar I, Pezer R, Zinic SD: **ColorHOR--novel graphical algorithm for fast scan of alpha satellite higher-order repeats and HOR annotation for GenBank sequence of human genome.** *Bioinformatics* 2005, **21**(7):846-852.
41. Gluncic M, Vlahovic I, Mrsic L, Paar V: **Global Repeat Map (GRM) Application: Finding All DNA Tandem Repeat Units.** *Algorithms* 2022, **15**(12).
42. Gluncic M, Vlahovic I, Paar V: **Discovery of 33mer in chromosome 21-the largest alpha satellite higher order repeat unit among all human somatic chromosomes.** *Sci Rep-Uk* 2019, **9**.
43. Gluncic M, Vlahovic I, Rosandic M, Paar V: **Tandemly repeated NBPF HOR copies (Olduvai triplets): Possible impact on human brain evolution.** *Life Sci Alliance* 2023, **6**(1).
44. Gluncic M, Vlahovic I, Rosandic M, Paar V: **Tandem NBPF 3mer HORs (Olduvai triplets) in Neanderthal and two novel HOR tandem arrays in human chromosome 1 T2T-CHM13 assembly.** *Sci Rep* 2023, **13**(1):14420.
45. Paar V, Gluncic M, Basar I, Rosandic M, Paar P, Cvitkovic M: **Large tandem, higher order repeats and regularly dispersed repeat units contribute substantially to divergence between human and chimpanzee Y chromosomes.** *J Mol Evol* 2011, **72**(1):34-55.
46. Paar V, Pavin N, Basar I, Rosandic M, Gluncic M, Paar N: **Hierarchical structure of cascade of primary and secondary periodicities in Fourier power spectrum of alphoid higher order repeats.** *BMC Bioinformatics* 2008, **9**:466.
47. Vlahović I, Glunčić M, Dekanić K, Mršić L, Jerković H, Martinjak I, V. P: **Global repeat map algorithm (GRM) reveals differences in alpha satellite number of tandem and higher order repeats (HORs) in human, Neanderthal and chimpanzee genomes – novel tandem repeat database.** *43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia* 2020:237-242.

48. Vlahovic I, Gluncic M, Rosandic M, Ugarkovic E, Paar V: **Regular Higher Order Repeat Structures in Beetle *Tribolium castaneum* Genome.** *Genome Biol Evol* 2017, **9**(10):2668-2680.
49. Rosandic M, Paar V, Gluncic M: **Fundamental role of start/stop regulators in whole DNA and new trinucleotide classification.** *Gene* 2013, **531**(2):184-190.
50. Wlodzimierz P, Hong M, Henderson IR: **TRASH: Tandem Repeat Annotation and Structural Hierarchy.** *Bioinformatics* 2023, **39**(5).
51. Smit AFA, Hubley R, Green P: **RepeatMasker Open-3.0. 1996–2010.**
52. Novak P, Neumann P, Macas J: **Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data.** *BMC Bioinformatics* 2010, **11**:378.
53. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**(2):573-580.
54. Kunyavskaya O, Dvorkina T, Bzikadze AV, Alexandrov IA, Pevzner PA: **Automated annotation of human centromeres with HORmon.** *Genome Res* 2022, **32**(6):1137-1151.
55. Bzikadze AV, Pevzner PA: **Automated assembly of centromeres from ultra-long error-prone reads.** *Nat Biotechnol* 2020, **38**(11):1309-1316.
56. Sevim V, Bashir A, Chin CS, Miga KH: **Alpha-CENTAURI: assessing novel centromeric repeat sequence variation with long read sequencing.** *Bioinformatics* 2016, **32**(13):1921-1924.
57. Gao S, Yang X, Guo H, Zhao X, Wang B, Ye K: **HiCAT: a tool for automatic annotation of centromere structure.** *Genome Biol* 2023, **24**(1):58.
58. Dvorkina T, Kunyavskaya O, Bzikadze AV, Alexandrov I, Pevzner PA: **CentromereArchitect: inference and analysis of the architecture of centromeres.** *Bioinformatics* 2021, **37**(Suppl_1):i196-i204.